DramaQA: Character-Centered Video Story Understanding with Hierarchical QA

Seongho Choi[1], Kyoung-Woon On[1], Yu-Jung Heo[1], Ahjeong Seo[1], Youwon Jang[1], Minsu Lee[1], Byoung-Tak Zhang[1, 2]
[1]Seoul National University, [2]AI Institute

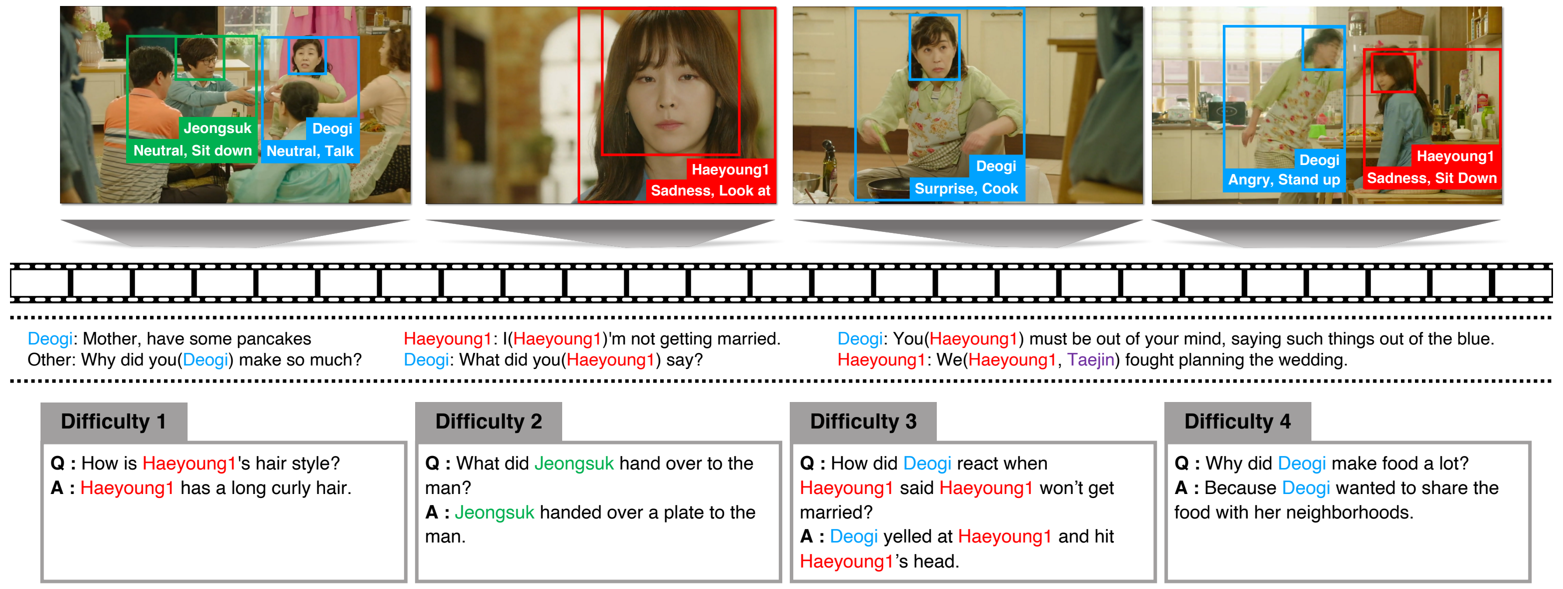
Background

- How to develop video story understanding models
  - One effective way is to train the models to answer questions about the video story.
  - e.g. TGIF-QA, MarioQA, PororoQA, MovieQA, TVQA

- How to evaluate the degree of intelligence of the models
  - The previous studies are highly-biased and lack of variance in the levels of question difficulty.

- Researches on how to evaluate the degree of video understanding based on human cognitive process have not progressed as yet.

DramaQA dataset



Deogi: Mother, have some pancakes
Other: Why did you(Deogi) make so much?

Haeyoung1: I(Haeyoung1)'m not getting married.
Deogi: What did you(Haeyoung1) say?

Deogi: You(Haeyoung1) must be out of your mind, saying such things out of the blue.
Haeyoung1: We(Haeyoung1, Taejin) fought planning the wedding.

Difficulty 1
Q : How is Haeyoung1's hair style?
A : Haeyoung1 has a long curly hair.

Difficulty 2
Q : What did Jeongsuk hand over to the man?
A : Jeongsuk handed over a plate to the man.

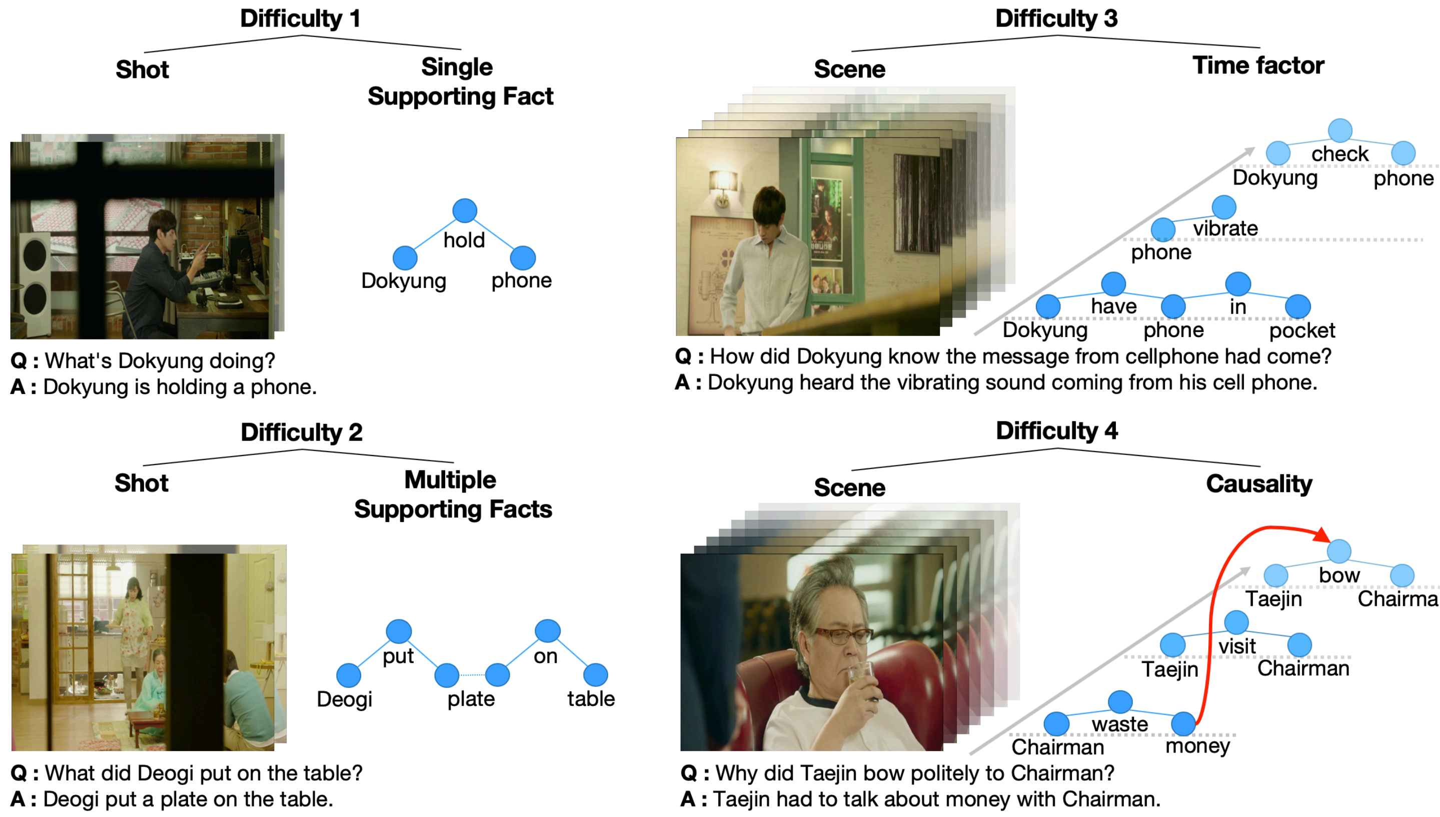
Difficulty 3
Q : How did Deogi react when Haeyoung1 said Haeyoung1 won't get married?
A : Deogi yelled at Haeyoung1 and hit Haeyoung1's head.

Difficulty 4
Q : Why did Deogi make food a lot?
A : Because Deogi wanted to share the food with her neighborhoods.

- Hierarchical QAs as an evaluation metric
  - Memory Capacity
  - Logical Complexity
- Character-centered video annotations
  - Visual metadata
    - bounding boxes, behaviors, and emotions of main characters
  - Coreferenced resolved scripts

Question-Answer Hierarchy

- Two criteria for classifying QAs into hierarchical levels of understanding
  - Memory Capacity is the required length of the video clip to answer
  - Logical Complexity is the required logical reasoning steps to answer



Q : What's Dokyung doing?
A : Dokyung is holding a phone.

Q : How did Dokyung know the message from cellphone had come?
A : Dokyung heard the vibrating sound coming from his cell phone.

Q : What did Deogi put on the table?
A : Deogi put a plate on the table.

Q : Why did Taejin bow politely to Chairman?
A : Taejin had to talk about money with Chairman.

Comparison with Other Datasets

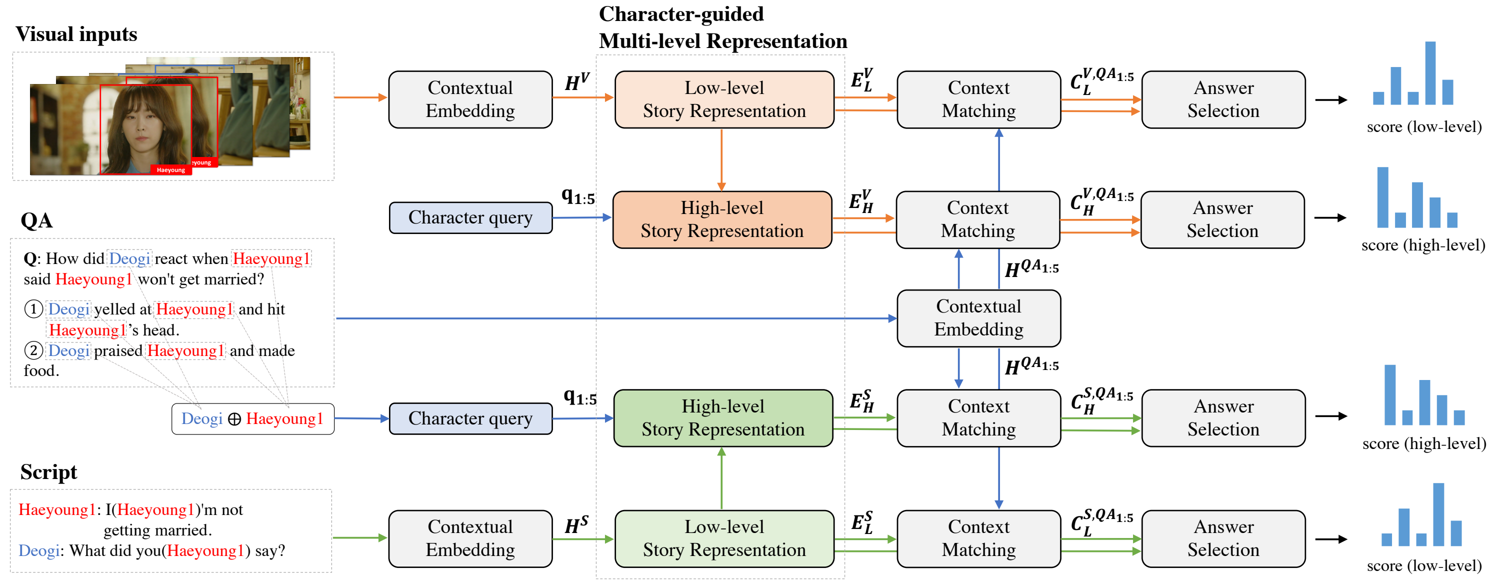
	# Q	# Annotated Images	Avg. Video len. (s)	Textual metadata	Visual metadata	Q. lev
TGIF-QA (Jang et al. 2017)	165,165	-	3.1	-	-	-
MarioQA (Mun et al. 2017)	187,757	-	< 6	-	-	-
PororoQA (Kim et al. 2017)	8,913	-	1.4	Description, Subtitle	-	-
MovieQA (Tapaswi et al. 2016)	6,462	-	202.7	Plot, DVS, Subtitle	-	-
TVQA (Lei et al. 2018)	152,545	-	76.2	Script	-	-
TVQA+ (Lei et al. 2019)	29,383	148,468	61.49	Script	Char./Obj. Bbox**	
DramaQA	17,983	217,308	3.7[a] 91.3[b]	Script*	Char. Bbox, Behavior, Emotion	✓

[a] Avg. video length for shot [b] Avg. video length for scene * Coreference resolved script ** Only mentioned in QAs

- DramaQA provides
  1) difficulty levels of the questions.
  2) annotations including visual metadata and coreference resolved scripts.
  3) tackles both shot-level and scene-level video clips.

Model and Ablation Study

- Overview of Multi-level Character Matching model



- Ablation Study

Model	Diff. 1	Diff. 2	Diff. 3	Diff. 4	Overall	Diff. Avg.
QA Similarity	30.64	27.20	26.16	22.25	28.27	26.56
S.Only−Coref	54.43	51.19	49.71	52.89	52.89	52.06
S.Only	62.03	63.58	56.15	55.58	60.95	59.34
V.Only−V.Meta	63.28	56.86	49.88	54.44	59.06	56.11
V.Only	74.82	70.61	54.60	56.48	69.22	64.13
Our−High	75.68	72.53	54.52	55.66	70.03	64.60
Our−Low	74.49	72.37	55.26	56.89	69.60	64.75
Our (Full)	75.96	74.65	57.36	56.63	71.14	66.15

Conclusion and Future Work

- The application area of the DramaQA dataset
  - emotion or behavior analysis of characters
  - automatic coreference identification from scripts
  - coreference resolution for visual-linguistic domain
  - action/face/object recognition or detection

- Future work of DramaQA dataset
  - extend the two criteria of hierarchical QA
  - provide hierarchical character-centered story descriptions
  - provide richer visual metadata including objects and places.

35th AAAI Conference on Artificial Intelligence (AAAI-21) | A Virtual Conference | February 2—9, 2021